排序下 PPS 抽样估计量的修正与应用

王 峰

（山西财经大学统计学院，山西 太原，030006）

摘要：受许多事物具有齐夫现象的启发，本文提出了排序后 PPS 抽样方法，并给出了修正汉森-赫维茨估计量及其方差。在此过程中本文解决了，长期以来抽样调查实践中将重要单元直接入样时，多少重要单元直接入样没有明确方法的问题，本文给出了理论依据和具体的确定方法。最后通过一个例子和中国城市人口抽样调查的案例，展示了修正汉森-赫维茨估计量的优势，并对这一研究方法做了总结和展望。

关键词：抽样调查；PPS 抽样；齐夫现象

中图分类号：C811　　　**文献标识码**：A

Modification and application of PPS sampling estimator under the order of rank

WANG Feng

(School of statistics, Shanxi University of Finance and Economics ,Shanxi, Taiyuan, 030006)

Abstract: Inspired by many things with Zipf phenomenon, this paper presents a modified Hansen-Hurwitz estimator and its variance. In the process, this paper solved the problem that sampling investigation practice for a long time will be the most important unit directly into the sample and how much the most important unit directly into the sample is no clear method. This paper gives the theoretical basis and specific method for determining. Finally, through an example and a case study of China's urban population sampling survey, the advantage of the modified Hansen-Hurwitz estimator was demonstrated, and the research method was summarized and forecasted.

Key words: sampling survey; probability proportional to size; Zipf phenomenon

1 引言

在抽样调查的实践中，多数情况下抽样单元的规模是不同的，因此各单元在总体中的地位也就不同。人们利用这种不同，让重要的单元入样概率高些，不重

基金项目：本文系国家统计局全国统计科学研究项目（2014LY013）的研究成果之一。

通讯作者：王峰，wangfeng@sxufe.edu.cn

要的单元入样概率低些，这种重要性通常由抽样单元的规模来衡量。由此人们提出了,每个单元在每次抽样中的概率与其单元的规模大小成比例，这种放回的与规模大小成比例的概率抽样就是 PPS（probability proportional to size）抽样。PPS 抽样在抽样调查实践中有广泛的应用，并被许多学者持续关注，对其进行研究和扩展。在 Hansen M H, Hurwitz W N.（1943）提出 PPS 抽样的理论[1]后，Yates F, Grundy P M.（1953）研究了层内应用 PPS 抽样的情况[2]；Holmberg A.（1998）将 bootstrap 的方法应用于 PPS 抽样；Kim Y W, Kim Y, Han H E（2013）研究了二阶段 PPS 系统抽样下的方差估计[3]；Patel P A, Bhatt S. A（2016）提出了 PPS 抽样基于模型的方差估计[4]；等等。在国内，邹国华，冯士雍（1995）研究了 PPS 抽样方案在放回抽样方案中的可溶性[5]；孙山泽，姜涛（2002）研究了 PPS 样本的轮换抽样[6]；刘建平，陈光慧(2005)分不同情况讨论了 MPPS 下汉森-赫维茨估计量的扩展[7]；等等[8-11]。国内外的研究普遍认为，衡量抽样单元重要性的指标抽样单元的规模是设计 PPS 抽样重要的辅助变量。那么是否可以考虑把抽样单元依据其规模排序，然后再设计 PPS 抽样呢。

为什么要对抽样单元依据其规模排序？齐夫在 1935 年分析了英语里单词的相对频率，发现一些单词拥有很高的频率。他得出，对所有词汇单词的频数与其排序后序号的乘积保持一个常数，这被认为是齐夫定律[12]。它揭示出，较少的单元拥有较高的频率或者对总量有较大的贡献，人们称这种性质为齐夫现象。后经研究发现，除语言应用外，诸如城市人口、个人收入、粮食生产等等很多方面都具有齐夫现象[13, 14]。受此启发，既然很多事物都具有齐夫现象，那么就可以考虑把总体单元按照规模排序后分成两部分，一部分为“重要”单元（齐夫现象的观点），让其全部入样，一部分为剩余部分再按照传统的 PPS 抽样来获取样本，以此来提高抽样的估计精度。在抽样实践中，将一些认为重要的单元直接入样的做法也较为多见，但是到底应该多少个重要的单元直接入样，通常根据实际情况由抽样设计者确定，没有固定的算法。本文将给出“重要”单元样本量的确定方法，并修正 PPS 抽样估计量——汉森-赫维茨估计量。

文章剩余部分将首先讨论抽样总体分成两部分后，PPS 抽样的估计量及方差的计算。然后利用方差的比较得出来“重要”单元的个数，以及使用条件。接下来利用教材中的例子和实际的案例来展示排序下修正 PPS 估计量的优势。文章的最后一部分是对这一方法的总结评价。

2 排序下 PPS 抽样估计量的修正

首先对总体单元按照其规模变量排序，令 Y_1，Y_2，……，Y_N 为排序后一有

限总体。M_1，M_2，……，M_N 是其对应的规模变量，$M_0=\sum_{i=1}^{N}M_i$，有 $z_i=\frac{M_i}{M_0}$。由传统的 PPS 抽样方法可知总体总值的无偏估计量为：

$$\hat{Y}_{HH}=\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{z_i} \tag{1}$$

式（1）为熟知的汉森-赫维茨估计量。

方差为：

$$V(\hat{Y}_{HH})=\frac{1}{n}\sum_{i=1}^{N}z_i(\frac{Y_i}{z_i}-Y)^2 \tag{2}$$

若 n＞1，则式（2）的无偏估计为：

$$v(\hat{Y}_{HH})=\frac{1}{n(n-1)}\sum_{i=1}^{n}(\frac{y_i}{z_i}-\hat{Y}_{HH})^2 \tag{3}$$

根据齐夫现象的观点，把总体分成两部分，第一部分 G_1 单位数为 N_1，全部入样，因此有 $n_1=N_1$；第二部分 G_2 单位数为 N_2，从中按传统 PPS 抽样的方法抽取 n_2，显然 $n_2=n-n_1$。为便于说明，令：

$$p_1=\sum_{i=1}^{N_1}z_i,\quad Y_1=\sum_{i=1}^{N_1}y_i,\quad p_2=\sum_{i=N_1+1}^{N}z_i,\quad Y_2=\sum_{i=N_1+1}^{N}y_i$$

把 G_2 看成一个总体的话，由 Hansen M H, Hurwitz W N.（1943）可以得到 G_2 总值的无偏估计量为：

$$\hat{Y}_{2HH}=\frac{1}{n_2}\sum_{i=1,i\in G_2}^{n_2}\frac{y_i}{z_i}$$

由于 G_1 是全部入样，不存在随机抽样。G_2 是 PPS 抽样，两部分合起来对总体总值的无偏估计量为：

$$\hat{Y}_{mHH}=\sum_{i=1}^{n_1}y_i+\frac{1}{n_2}\sum_{i=1,i\in G_2}^{n_2}\frac{y_i}{z_i} \tag{4}$$

称式（4）为修正的汉森-赫维茨估计量。注意到，G_1 是全部入样，因此 $\sum_{i=1}^{n_1}y_i$ 不是一个随机变量，在式（4）中相当于一个常数，而 $\frac{1}{n_2}\sum_{i=1,i\in G_2}^{n_2}\frac{y_i}{z_i}$ 是汉森-赫维

茨估计量，因此由汉森-赫维茨估计量的无偏性可以得出 $\frac{1}{n_2}\sum_{i=1,i\in G_2}^{n_2}\frac{y_i}{z_i}$ 是 G_2 总值的无偏估计。$\sum_{i=1}^{n_1} y_i$ 本身就是 G_1 的总值。故这两部分的和 $\hat{Y}_{mHH}$ 就是总体总值的无偏估计。

同样的道理，修正的汉森-赫维茨估计量第一部分为非随机变量，第二部分是实施 PPS 抽样的汉森-赫维茨估计量，因此修正的汉森-赫维茨估计量的方差为：

$$V(\hat{Y}_{mHH})=\frac{1}{n_2}\sum_{i=N_1+1}^{N}\frac{z_i}{p_2}(\frac{p_2Y_i}{z_i}-Y_2)^2 \quad (5)$$

同理，若 n_2 >1，可由 Hansen M H, Hurwitz W N.（1943）得到式（5）的无偏估计量为：

$$v(\hat{Y}_{mHH})=\frac{1}{n_2(n_2-1)}\sum_{i=1,i\in G_2}^{n_2}(\frac{y_i}{z_i}-\hat{Y}_{2HH})^2 \quad (6)$$

3 比较 pps 抽样和修正 pps 抽样总值估计量的方差

pps 抽样下汉森-赫维茨估计量的方差式（2）可以变形为：

$$V(\hat{Y}_{HH})=\frac{1}{n}\sum_{i=1}^{N_1}z_i(\frac{Y_i}{z_i}-Y)^2+\frac{1}{n}\sum_{i=N_1+1}^{N}z_i(\frac{Y_i}{z_i}-Y)^2$$

对于修正的汉森-赫维茨估计量的方差有：

$$V(\hat{Y}_{mHH})=\frac{1}{n_2}\sum_{i=N_1+1}^{N}\frac{z_i}{p_2}(\frac{p_2Y_i}{z_i}-Y_2)^2$$

$$=\frac{p_2}{n_2}\sum_{i=N_1+1}^{N}z_i(\frac{Y_i}{z_i}-Y+Y-\frac{Y_2}{p_2})^2$$

$$=\frac{p_2}{n_2}\left(\sum_{i=N_1+1}^{N}z_i(\frac{Y_i}{z_i}-Y)^2+\sum_{i=N_1+1}^{N}z_i(Y-\frac{Y_2}{p_2})^2+2\sum_{i=N_1+1}^{N}z_i(\frac{Y_i}{z_i}-Y)(Y-\frac{Y_2}{p_2})\right)$$

$$=\frac{p_2}{n_2}\left(\sum_{i=N_1+1}^{N}z_i(\frac{Y_i}{z_i}-Y)^2+(Y-\frac{Y_2}{p_2})^2\sum_{i=N_1+1}^{N}z_i+2(Y-\frac{Y_2}{p_2})\sum_{i=N_1+1}^{N}z_i(\frac{Y_i}{z_i}-Y)\right)$$

$$=\frac{p_2}{n_2}\left(\sum_{i=N_1+1}^{N}z_i(\frac{Y_i}{z_i}-Y)^2+(Y-\frac{Y_2}{p_2})^2p_2-2(Y-\frac{Y_2}{p_2})^2p_2\right)$$

$$=\frac{p_2}{n_2}\left(\sum_{i=N_1+1}^{N} z_i(\frac{Y_i}{z_i}-Y)^2-(Y-\frac{Y_2}{p_2})^2 p_2\right)$$

两方差相减得：

$$V(\hat{Y}_{HH})-V(\hat{Y}_{mHH})$$

$$=\frac{1}{n}\sum_{i=1}^{N_1} z_i(\frac{Y_i}{z_i}-Y)^2+\frac{1}{n}\sum_{i=N_1+1}^{N} z_i(\frac{Y_i}{z_i}-Y)^2$$

$$-\frac{p_2}{n_2}\left(\sum_{i=N_1+1}^{N} z_i(\frac{Y_i}{z_i}-Y)^2-(Y-\frac{Y_2}{p_2})^2 p_2\right)$$

$$=\frac{1}{n}\sum_{i=1}^{N_1} z_i(\frac{Y_i}{z_i}-Y)^2+(\frac{1}{n}-\frac{p_2}{n_2})\sum_{i=N_1+1}^{N} z_i(\frac{Y_i}{z_i}-Y)^2+\frac{p_2^2}{n_2}(Y-\frac{Y_2}{p_2})^2 \quad (7)$$

很显然，上述式（7）由三部分构成，第一部分和第三部分均非负。如果第二部分中$\frac{1}{n}-\frac{p_2}{n_2}>0$的话，则有$V(\hat{Y}_{HH})>V(\hat{Y}_{mHH})$。此时修正的汉森-赫维茨估计量就会优于传统的汉森-赫维茨估计量。

4 “重要”单元个数$n_1=N_1$的确定

要找到“重要”的单元数n_1，其实就是找到满足$\frac{1}{n}-\frac{p_2}{n_2}>0$的$n_1$。因为$n_2=n-n_1$，$p_2=1-p_1$，就有：

$$\frac{1}{n}-\frac{p_2}{n_2}>0 \Leftrightarrow \frac{1}{n}>\frac{1-p_1}{n-n_1} \Leftrightarrow n-n_1>n-np_1 \Leftrightarrow np_1>n_1$$

由上述不等式知，如果p_1很小，使得np_1都小于 1 了，显然这样的n_1无法得到，因为n_1不可能小于 1，也就不能用修正的汉森-赫维茨估计量，这种情况下可用传统的汉森-赫维茨估计量来做。另一方面也可以通过扩大样本量，使其np_1>1，这一点在后面的例子中有详细说明。如果np_1>1 了，就可以使用修正的汉森-赫维茨估计量。

接下来确定n_1。由上分析知np_1>1，否则无法修正汉森-赫维茨估计量。而

修正汉森-赫维茨估计量的条件为$\frac{1}{n}-\frac{p_2}{n_2}>0$，即$n_1<np_1 \Leftrightarrow \frac{p_2}{n_2}<\frac{1}{n}$；根据这个不等式由抽样原理可知，当$\frac{p_2}{n_2}$取最小值时$\frac{1}{n}-\frac{p_2}{n_2}$达到最大值，此时一定有$V(\hat{Y}_{HH})>V(\hat{Y}_{mHH})$。因此就需要在所有可能的$n_1=1,2,3,\cdots,n-1$中找到使$\frac{p_2}{n_2}$达到最小值时的$n_1$，令此时的$n_1=n_j$。也就是说，当$n_1=1$时，由条件知$np_1>1$，否则无法进行后续计算，理由如前所述，这里不再赘述。由于$n_1<np_1 \Leftrightarrow \frac{p_2}{n_2}<\frac{1}{n}$，因此当$n_1=1$时，必须满足$\frac{p_2}{n_2}<\frac{1}{n}$。当$n_1=2,3,\cdots$按自然数递增时，$n_2$则是按照自然数递减的，即$n_2=(n-2),(n-3),\cdots$。注意到，总体单位是按照规模排序的，一定有$z_1>z_2>\cdots\cdots>z_n$，且$z_i<1$，$i=1,2,\cdots,n$，由$p_2=1-z_1-z_2,1-z_1-z_2-z_3,\cdots$，所以$p_2$先递减的快后递减的慢，且按$z_i$（$z_i<1$）递减。由此$\frac{p_2}{n_2}$通常会表现为先减小后增加。（具体的数学证明见附录）所以在$n_1=n_j$时，$\frac{p_2}{n_2}$达到最小值，此时也一定满足$\frac{p_2}{n_2}<\frac{1}{n}$。这时候必然有$V(\hat{Y}_{HH})>V(\hat{Y}_{mHH})$成立，即修正的汉森-赫维茨估计量就会优于传统的汉森-赫维茨估计量。如果$\frac{p_2}{n_2}$一直在增加，那就说明$n_j=n_1=1$时，$\frac{p_2}{n_2}$为最小。如果连$n_1=1$都无法满足$\frac{p_2}{n_2}<\frac{1}{n}$，也就是未能满足初始条件，那这个数据就不能用修正的汉森-赫维茨估计量，这时候就用传统的方法即可。接下来用实例验证修正的汉森-赫维茨估计量。

5 实证分析与检验

这部分用两个例子来说明修正汉森-赫维茨估计量的应用，并与汉森-赫维茨估计量做比较。

5.1 来自教材的一个实例

这个例子来自于教材，冯士雍，倪加勋，邹国华（1998）中的例 7.2 职工人数调查[15]。该例子陈述如下：

表 1 为某系统全部 N=36 个单位上一年职工人数 X_i 及当年职工人数 Y_i 的数据。以 X_i 为单位大小 M_i 的度量，对单位进行 PPS 抽样，$n=11$，估计全系统当年职工总人数 Y。

表 1 某系统各单位上一年与当年职工人数

单位号	X	Y	单位号	X	Y	单位号	X	Y
1	598	633	13	497	516	25	252	281
2	21	18	14	723	786	26	194	210
3	630	656	15	712	740	27	149	166
4	3012	3273	16	335	352	28	173	189
5	372	386	17	267	299	29	318	344
6	142	164	18	1658	1714	30	204	227
7	1072	1145	19	231	255	31	52	63
8	432	501	20	15	24	32	188	174
9	216	235	21	172	181	33	97	122
10	1698	1778	22	234	243	34	218	242
11	1570	1541	23	312	338	35	47	51
12	502	486	24	351	371	36	838	879

资料来源：冯士雍，倪加勋，邹国华．抽样调查理论与方法．中国统计出版社，1998．

第一步：依据规模变量 X_i 从大到小排序，计算 np_1，并判断是否满足修正汉森-赫维茨估计量的条件。

计算得 p_1=0.163，n=11；满足 np_1＞1。这里说明一点，教材中原例题取 6 个样本，如果 n=6，即出现了 np_1＜1 的情况无法满足使用修正汉森-赫维茨估计量的条件。一种可以使用教材中的方法，利用汉森-赫维茨估计量去估计；另一种就是扩大样本量，使其满足 np_1＞1，从而可以使用修正汉森-赫维茨估计量。由于其他教材中的例子多为只给出 PPS 样本而没有给出抽取的总体数据，从而没办法抽样比较。因此这里选用该例子并把抽取样本数改为 11，这并不影响与汉森-赫维茨估计量的比较。

第二步：确定 n_1。

由 $p_1=\sum_{i=1}^{N_1} z_i$，$N_1=n_1$，计算出所有可能 N_1 下的 p_1，$p_2=1-p_1$，由此计算出 p_2/n_2，并找出使 p_2/n_2 取最小值的 n_1，计算结果见表 2

表 2　　　　　　　　n_1 的确定

n1	n2	p1	p2	p2/n2
1	10	0.163102	0.836898	0.08369
2	9	0.25505	0.74495	0.082772
3	8	0.344831	0.655169	0.081896
4	**7**	**0.42985**	**0.57015**	**0.081450**
5	6	0.487897	0.512103	0.085350
6	5	0.533276	0.466724	0.093345
7	4	0.572426	0.427574	0.106893
8	3	0.610982	0.389018	0.129673
9	2	0.645097	0.354903	0.177452
10	1	0.677479	0.322521	-

由图 1 可以看出 n_1=4 时 p_2/n_2 为最小，即“重要”单元的个数 $N_1=n_1$ 确定完成，也就是排在前 4 个的单元为 G_1 全部入样。单位号分别为：4，10，18，11。

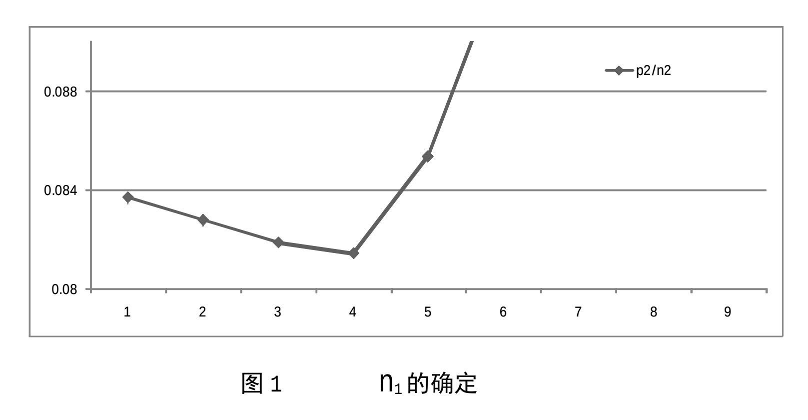


图 1　　　　n_1 的确定

第三步：去除总体中前 n_1 个单元后，在剩余的总体中按 PPS 抽样的方法抽取 $n_2=n-n_1$ 个单元，至此抽取样本完成。

PPS 抽样的方法很多，这里用代码法抽取 n_2 个单元，得到 G_2 中的样本单位号为 17，25，12，23，36，15，3。

接下来就可以用抽取的样本，利用式（4）算出修正汉森-赫维茨估计量，估计该系统当年的职工总人数，式（6）可以估计出该估计量的方差。与冯士雍，倪加勋，邹国华（1998）类似，这里也用如上三步同样的方法抽取出四组样本，单位号分别为：

样本I：4，10，18，11；17，25，12，23，36，15，3

样本II：4，10，18，11；24，14，19，21，7，36，25

样本III：4，10，18，11；5，24，14，1，17，3，13

样本IV：4，10，18，11；13，24，15，25，19，21，1

计算出修正汉森-赫维茨估计量和汉森-赫维茨估计量及各自估计量方差的估计值，见表 3

表 3　两种方法的总值估计及标准差的比较

样本	$\hat{Y}_{mHH}$	$\sqrt{v(\hat{Y}_{mHH})}$	$\hat{Y}_{HH}$	$\sqrt{v(\hat{Y}_{HH})}$	$\sqrt{V(Y)}$
样本 I	19791.99	335.7994	19780.62	411.8905	647.8042
样本 II	19671.38	104.9526	19646.18	205.9549	647.8042
样本III	19531.34	121.9562	19490.09	200.4539	647.8042
样本IV	19568.66	122.5135	19531.69	204.6205	647.8042

表 3 第二列和第三列分别是修正汉森-赫维茨估计量及估计量的标准差，第四列和第五列分别是同样本下传统汉森-赫维茨估计量及估计量的标准差。从中可以看出，修正后的汉森-赫维茨估计量的标准差均小于其相应汉森-赫维茨估计量的标准差。这一结果与前述理论一致，修正后的汉森-赫维茨估计量要优于汉森-赫维茨估计量。另外需要说明的是，冯士雍，倪加勋，邹国华（1998）中汉森-赫维茨估计量的方差与本文中汉森-赫维茨估计量的方差不一样，多数偏大。一方面是因为抽选的样本不同，更重要的是本文的样本量（n=11）比前者样本量（n=6）大。但是，同样本情况下相比，修正汉森-赫维茨估计量要明显优于汉森-赫维茨估计量。

5.2 中国城市人口抽样调查

接下来通过对中国 655 个城市人口调查来说明修正汉森-赫维茨估计量的应用及优势。以最近一次的人口普查数据（2010 年）为依据，对中国 655 个城市依据其人口数从大到小排序。估计中国 655 个城市的人口数，这里取样本量约为总数的 10%，即为 66 个城市来估计调查年份（2014 年）中国 655 个城市的总人口。数据来自于 http://www.stats.gov.cn/。

第一步：按照普查年份的人口数对 655 个城市排序，计算 np_1=1.62 大于 1，可以使用修正汉森-赫维茨估计量。

第二步：计算所有可能 n_1 下的 p_2/n_2 的值，并找出使 p_2/n_2 取最小值的 n_1。这里只列出了前 10 个，后面 p_2/n_2 依次增加故省略。

表 4　n_1 的确定

n1	n2	p1	p2	p2/n2
1	65	0.024488	0.975512	0.015008

2	64	0.045565	0.954435	0.014913
3	**63**	**0.064251**	**0.935749**	**0.01485**
4	62	0.077379	0.922621	0.014881
5	61	0.090113	0.909887	0.014916
6	60	0.100536	0.899464	0.014991
7	59	0.109364	0.890636	0.015096
8	58	0.117967	0.882033	0.015207
9	57	0.126364	0.873636	0.015327
10	56	0.134471	0.865529	0.015456
…	…	…	…	…

由表 4 可以看出 n_1=3 时，p_2/n_2 达到最小值。这 3 个“重要”单元的城市分别是排序后排在前三位的重庆市，上海市，北京市。

第三步，用 PPS 抽样的方法抽取剩下的 63 个样本。得到 66 个样本城市见表 5。

已知普查年份城市人口数为 637359998 人，根据式（4）得到调查年份城市人口数为 659772963 人，式（6）得到估计量标准差的估计值是 2996092 人。如果用汉森-赫维茨估计量得到调查年份城市人口数为 658451503 人，估计量标准差的估计值是 3710747 人，要比修正后的汉森-赫维茨估计量标准差大很多。如果按照冯士雍，倪加勋，邹国华（1998）中重复多次上述操作比较其结果，也能得出上述类似结论。

6 总结评价与展望

通过上述的论证和实例的检验与分析可以看出，修正汉森-赫维茨估计量要明显优于传统的汉森-赫维茨估计量。而齐夫现象的广泛性，也说明了本文所提方法应用范围的广泛。当然，如果在应用中没有满足修正汉森-赫维茨估计量的使用条件即 $np_1 > 1$，可以通过扩大样本量使其满足条件，然后再应用修正汉森-赫维茨估计量。值得一提的是，本文还解决了长期以来在抽样的实践中将部分“重要”单元直接入样，到底多少“重要”单元入样为宜没有明确方法的问题。本文给出了明确的方法和计算公式，即在所有可能的 n_1 中，找使 p_2/n_2 取最小值的 n_1，这个 n_1 就是需要直接入样的单元数。本文最后的两个案例再次印证了修正汉森-赫维茨估计量应用的广泛性和优势。可以预见，将已证明的事物规律应用于抽样调查，不会仅限于此，更多的辅助信息与方法会更好的应用于抽样调查，促进抽样调查水平的不断提高。

表 5　　66 个样本城市普查年份和调查年份人口数　　单位：人

城市	2010 年	2014 年	城市	2010 年	2014 年	城市	2010 年	2014 年

重庆市	**15607433**	**19438702**	菏泽市	1527571	1555668	双城市	821756	808283
上海市	**13433709**	**13709152**	遂宁市	1504196	1523315	普兰店市	817837	923145
北京市	**11909663**	**12631493**	宣威市	1478563	1529228	铜川市	759827	747664
武汉市	8367323	8273117	简阳市	1468168	1487044	义乌市	739838	766604
天津市	8116493	8344259	齐齐哈尔市	1415146	1382338	盖州市	728823	703959
广州市	6642840	6949637	漯河市	1408206	1341481	舟山市	697226	709040
西安市	5626490	5871627	大庆市	1333657	1365550	攀枝花市	689654	683575
沈阳市	5154241	5284407	榆树市	1304436	1275220	长乐市	685105	715790
杭州市	4348166	4584653	福清市	1275016	1335158	荣成市	670251	669403
长春市	3627536	3658620	泰兴市	1196164	1198791	清远市	655672	672375
乌鲁木齐市	2335780	2606434	瑞安市	1190519	1231071	涿州市	645542	670571
普宁市	2325688	2444622	吴川市	1101691	1178370	石河子市	629651	637204
合肥市	2155767	2453691	资阳市	1089424	1105495	西昌市	618540	652947
兰州市	2103639	2048802	晋江市	1065770	1108142	霸州市	618273	639732
福州市	1885939	1974319	章丘市	1015129	1023903	贵溪市	600398	637941
南阳市	1885076	1974549	肥城市	978866	988737	晋中市	595208	612206
六安市	1865174	1891181	大冶市	942641	958960	平湖市	486996	491379
陆丰市	1770654	1886043	金华市	931854	950886	南宫市	476096	497387
商丘市	1769870	1804862	庄河市	905852	903662	伊宁市	471462	559691
滕州市	1681431	1693074	保山市	900024	925523	延安市	458166	464885
化州市	1612431	1698541	岑溪市	898166	933674	铜陵市	448284	448738
宿迁市	1597733	1720004	利川市	895597	917101	中卫市	395899	406426

[参考文献]

[1] Hansen M H, Hurwitz W N. On the Theory of Sampling from Finite Populations[J]. Annals of the Rheumatic Diseases. 1943, 70(12): 2111-2118.

[2] Yates F, Grundy P M. Selection without replacement from within strata with probability proportional to sue[J]. Journal of the Royal Statistical Society. 1953, 15(2): 253-261.

[3] Kim Y, Kim Y, Han H, et al. Efficiency of Variance Estimators for Two-stage PPS Systematic Sampling[J]. Korean Journal of Applied Statistics. 2013, 26(6): 1033-1041.

[4] Patel P A, Bhatt S. A Model-based Estimation of Finite population Variance under PPS Sampling[J]. Imperial Journal of Interdisciplinary Research. 2016, 2(4).

[5] 邹国华，冯士雍．放回的PPS抽样方案在放回抽样方案类中的可容许性[J]．科学通报．1995(08): 683-686.

[6] 孙山泽，姜涛. PPS样本的轮换抽样[J]．数理统计与管理. 2002(04): 61-64.

[7] 刘建平，陈光慧. MPPS抽样下Hansen-Hurwitz估计量的扩展[J]．统计研究. 2005(05): 50-53.

[8] 陈光慧，曹伟伟．半参数乘积调整模型的抽样估计方法及应用研究[J]．数理统计与管理．2017: 1-14.

[9] 李莉莉．基于Brewer抽样的不放回样本追加策略下域的估计[J]．数理统计与管理．2017(04): 651-660.

[10] 孟令宾，李二倩，田茂再．基于鞍点逼近的二项抽样下优势比的置信区间构造[J]．数理统计与管理. 2017(01): 85-102.

[11] 米子川，李毅．面向SNS大数据的捕获移出模型抽样估计[J]．数理统计与管理．2016(03): 424-434.
[12] Zipf G. The Psycho-Biology of Language. An Introduction to Dynamic Philology[J]. Journal of Nervous & Mental Disease. 1935, 85(1): 93.
[13] 张忠友．齐夫定律的理论基础及其实践意义[J]．情报科学. 1989(5): 62-66.
[14] 徐兴余. 20/80律与布—齐—洛三个定律之间的关系[J]．图书情报工作. 2003(8): 39-42.
[15] 冯士雍，倪加勋，邹国华．抽样调查理论与方法[M]．中国统计出版社, 1998.

附录：

证明：$\dfrac{p_2}{n_2}$序列随n_1的增大将表现为先减小后增大，或者一直增大。

证明：为便于说明，令$n_1=1, 2, \cdots, i, \cdots$, n-1 生成的$\dfrac{p_2}{n_2}$数列为$a_1, a_2, \cdots, a_i, \cdots, a_{n-1}$

此时，

$n_2 = n-1, n-2, \cdots, n-i, \cdots, 1$

$p_1 = z_1, z_2, \cdots, z_i, \cdots z_{n-1}$；即$z_i$，$i = 1, 2, \cdots, n-1$，且$z_1 > z_2 > \cdots z_i > \cdots > z_{n-1}$，$0 < z_i < 1$

$p_2 = 1-z_1, 1-z_2, \cdots, 1-z_i, \cdots, 1-z_{n-1}$

当$n_1 = 1$时，由使用修正汉森-赫维茨估计量条件知$np_1 > n_1 = 1$，　$p_1 > \dfrac{1}{n}$，即

$z_1 > \dfrac{1}{n}$。此时有$a_1 = \dfrac{p_2}{n_2} = \dfrac{1-z_1}{n-1}$

当$n_1 = 2$时，$a_2 = \dfrac{p_2}{n_2} = \dfrac{1-z_1-z_2}{n-2}$

考察：$a_1 - a_2 = \dfrac{1-z_1}{n-1} - \dfrac{1-z_1-z_2}{n-2} = \dfrac{1-z_1}{n-1} - \dfrac{1-z_1}{n-2} + \dfrac{z_2}{n-2} = \left(\dfrac{-1}{(n-1)(n-2)}\right)(1-z_1) + \dfrac{z_2}{n-2}$

若要使$a_1 > a_2$，只需要$\left(\dfrac{-1}{(n-1)(n-2)}\right)(1-z_1) + \dfrac{z_2}{n-2} > 0$

即$z_2 > \dfrac{1}{n-1}(1-z_1)$，$\because z_1 > \dfrac{1}{n}$，$\therefore 1-z_1 < 1-\dfrac{1}{n}$

于是只需要使$z_2 > \dfrac{1}{n-1}\dfrac{n-1}{n} = \dfrac{1}{n}$ 即可，

所以只要 $z_2 > \frac{1}{n}$，就有 $a_1 > a_2$。

同理，

$$a_i - a_{i+1} = \frac{1 - z_1 - z_2 - \cdots - z_i}{n-i} - \frac{1 - z_1 - z_2 - \cdots - z_{i+1}}{n-(i+1)}$$

$$= \frac{1 - z_1 - z_2 - \cdots - z_i}{n-i} - \frac{1 - z_1 - z_2 - \cdots - z_i}{n-(i+1)} + \frac{z_{i+1}}{n-(i+1)}$$

$$= \left(\frac{-1}{(n-i)\left[n-(i+1)\right]} \right)(1 - z_1 - z_2 - \cdots - z_i) + \frac{z_{i+1}}{n-(i+1)}$$

若要使 $a_i > a_{i+1}$，需要

$$\left(\frac{-1}{(n-i)\left[n-(i+1)\right]} \right)(1 - z_1 - z_2 - \cdots - z_i) + \frac{z_{i+1}}{n-(i+1)} > 0$$

即 $z_{i+1} > \frac{1}{n-i}(1 - z_1 - z_2 - \cdots - z_i)$，

$\because z_1, z_2, \cdots, z_i > \frac{1}{n}$，$\therefore 1 - z_1 - z_2 - \cdots - z_i < 1 - \frac{i}{n} = \frac{n-i}{n}$

于是只需要使 $z_{i+1} > \frac{1}{n-i} \frac{n-i}{n} = \frac{1}{n}$ 即可，

因此序列 a_i，即 $\frac{p_2}{n_2}$ 随 n_1 的增大在减小。

注意到：$z_1 > z_2 > \cdots z_i > \cdots > z_{n-1}$，且 $0 < z_i < 1$，一定会有 $z_j < \frac{1}{n}$

此时，$\frac{p_2}{n_2}$ 开始增大。

综上，$\frac{p_2}{n_2}$ 先减小后增大。

如果 $z_1 < \frac{1}{n}$，那么 $\frac{p_2}{n_2}$ 就一直增大。